E5874

MENU SEARCH INDEX DETAIL

1 / 1

PATENT ABSTRACTS OF JAPAN

(11)Publication number : 06-139027
(43)Date of publication of application : 20.05.1994

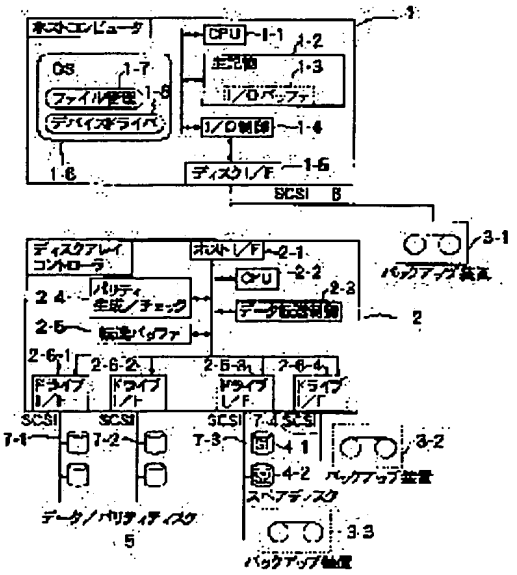
(51)Int.Cl. G06F 3/06
G06F 3/06
G11B 19/02

(21)Application number : 04-290428
(22)Date of filing : 28.10.1992

(71)Applicant : HITACHI LTD
(72)Inventor : OEDA TAKASHI
YOSHIDA MINORU
HONDA KIYOSHI
MATSUNAMI NAOTO
MIYAZAWA SHOICHI
ISONO SOICHI

(54) DATA BACKUP METHOD FOR DISK DRIVER AND DISK ARRAY DEVICE AND DATA STORAGE SYSTEM AND DISK ARRAY SYSTEM

(57)Abstract:
PURPOSE: To shorten the backup occupying time of a disk array device.
CONSTITUTION: A data/parity disk drive 5 which stores data or parity information is connected like an array through drive I/F 2-6-1, 2-6-2,... with a disk array controller 2. And also, the prescribed number of spare disk drives 4 are connected with the drive I/F 2-6-3 of the disk array controller 2. When a fault occurs at the data/parity disk drive 5, the spare disk drive 4 can be used as the data/parity disk drive instead of the defective data parity/parity disk drive, or used as a data transfer buffer between the data/parity disk drive 5 and backup devices 3-1, 3-2, and 3-3 at the time of a data trasfer for backup- processing data on the disk array by backup devices.


CPU
OS
SCSI
SCSI
SCSI

JP#05

E5874

(19)日本国特許庁（ＪＰ）　(12)公開特許公報（Ａ）　(11)特許出願公開番号

特開平6－139027

(43)公開日　平成６年(1994)５月20日

(51)Int.Cl.[5]	識別記号	庁内整理番号	ＦＩ	技術表示箇所
Ｇ０６Ｆ　3/06	３０１　Ｚ	7165－5B		
	３０４　Ｆ	7165－5B		
Ｇ１１Ｂ　19/02		Ｆ　7525－5D		

審査請求　未請求　請求項の数20(全 22 頁)

(21)出願番号　特願平4－290428

(22)出願日　平成４年(1992)10月28日

(71)出願人　000005108
株式会社日立製作所
東京都千代田区神田駿河台四丁目６番地

(72)発明者　大枝　高
神奈川県横浜市戸塚区吉田町292番地　株式会社日立製作所マイクロエレクトロニクス機器開発研究所内

(72)発明者　吉田　稔
神奈川県小田原市国府津2880番地　株式会社日立製作所小田原工場内

(74)代理人　弁理士　武　顕次郎

最終頁に続く

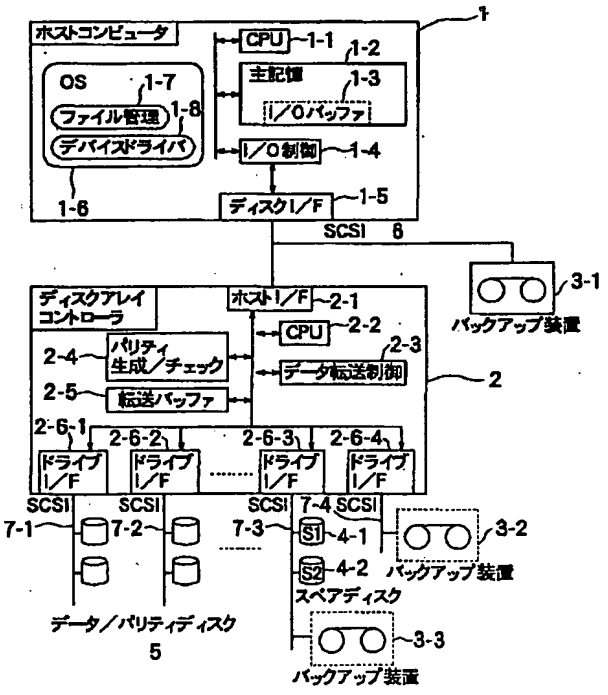
(54)【発明の名称】　ディスクドライバ，ディスクアレイ装置，データ記憶システム及びディスクアレイシステムのデータバックアップ方法

(57)【要約】

【目的】　ディスクアレイ装置のデータのバックアップ占有時間を短縮する。

【構成】　ディスクアレイコントローラ２に、ドライブＩ／Ｆ２－６－１，２－６－２，……を介し、データやパリティ情報を格納するデータ／パリティデイスクドライブ５がアレイ状に接続されており、また、このディスクアレイコントローラ２のドライブＩ／Ｆ２－６－３にスペアディスクドライブ４が所定個数接続されている。かかるスペアディスクドライブ４は、データ／パリティデイスクドライブ５に障害が生じたとき、この障害データ／パリティデイスクドライブに代ってデータ／パリティデイスクドライブとなるが、かかるディスクアレイ上でのデータをバックアップ装置３－１，３－２または３－３でバックアップのためのデータ転送に際し、データ／パリティデイスクドライブ５とかかるバックアップ装置との間のデータ転送バッファにもなる。

【図　1】



【特許請求の範囲】
【請求項１】　ディスクアレイコントローラと、ドライブインターフェースによって該ディスクアレイコントローラに接続され、アレイ状に配列された複数のデータ等を格納するためのディスクドライブと、該ディスクドライブに格納されているデータをバックアップするためのデータバックアップ装置とを備え、
該複数のディスクドライブのうちの１以上のディスクドライブを予備ディスクドライブとし、残りをデータ等を格納するデータ／パリティディスクドライブとし、該予備ディスクドライブを、障害が生じた該データ／パリティディスクドライブに代ってデータ／パリティディスクドライブとするディスクアレイ装置において、
データのバックアップの際、該予備ディスクドライブを該データ／パリティディスクドライブと該データバックアップ装置との間のデータ転送バッファとすることを特徴とするディスクアレイ装置。
【請求項２】　請求項１において、
前記データバックアップ装置を、前記ディスクアレイコントローラのドライブインターフェースに接続し、前記ディスクアレイコントローラからの制御を可能に構成したことを特徴とするディスクアレイ装置。
【請求項３】　請求項２において、
前記データ／パリティディスクドライブはインターフェースを２ポート以上有し、前記ディスクアレイコントローラと前記データバックアップ装置とに接続されていることを特徴とするディスクアレイ装置。
【請求項４】　請求項１において、
前記アレイ上に配列された前記複数のディスクドライブ上での現在のファイルシステムの使用範囲を認識する手段もしくは上位コンピュータから、その認識結果を表わす情報を獲得する手段を有し、
該上位コンピュータが介在することなく、該フアイルシステムの使用範囲のデータの前記データバックアップ装置へのバックアップの進行、及び前記アレイ上に配列された前記複数のディスクドライブ上への前記データバックアップ装置のバックアップデータの回復を行なうことを特徴とするディスクアレイ装置。
【請求項５】　請求項４において、
前回のバックアップ時から変更されたアドレスを認識する手段もしくは前記上位コンピュータから、その認識結果を表わす情報を獲得する手段を有し、
前記上位コンピュータが介在することなく、前記データバックアップ装置への差分バックアップの進行、及び前記アレイ上に配列された前記複数のディスクドライブ上への前記データバックアップ装置のバックアップデータの回復を行なうことを特徴とするディスクアレイ装置。
【請求項６】　請求項１において、
前記データバックアップ装置へのバックアップが実行されている前記アレイ上に配列された前記複数のディスクドライブ上でのバックアップ対象ファイルシステムへのアクセスを記録しておく手段を有し、
上位コンピュータから該バックアップ対象ファイルシステムへのアクセスをしながら、該バックアップ対象ファイルシステムのデータのバックアップができることを特徴とするディスクアレイ装置。
【請求項７】　１もしくは複数のインターフェースによってコンピュータシステムと接続され、かつ１もしくは複数のドライブインターフェース夫々に複数の記憶装置が接続されてなり、該複数の記憶装置によってフアイルデータを記憶できるようにしたデータ記憶システムにおいて、
該記憶装置の少なくとも１つは予備の記憶装置であって、
データバックアップ手段と、
データを記憶する記憶装置に障害が発生したとき、該障害が発生した記憶装置に代えて、その論理的位置を該予備の記憶装置で占めさせる手段と、
該記憶装置のデータのデータバックアップ時、該予備の記憶装置を該記憶装置と該バックアップ手段との間のデータ転送バッファとする手段とを有することを特徴とするデータ記憶システム。
【請求項８】　請求項７において、
前記複数の記憶装置上での現在のファイルシステムの使用範囲を認識する手段もしくは上位コンピュータから、その認識結果を表わす情報を獲得する手段を有し、
該上位コンピュータが介在することなく、該フアイルシステムの使用範囲のデータの前記データバックアップ手段へのバックアップの進行、及び前記複数の記憶装置上への前記データバックアップ手段のバックアップデータの回復を行なうことを特徴とするデータ記憶システム。
【請求項９】　請求項８において、
前回のバックアップ時から変更されたアドレスを認識する手段もしくは前記上位コンピュータから、その認識結果を表わす情報を獲得する手段を有し、
前記上位コンピュータが介在することなく、前記データバックアップ手段への差分バックアップの進行、及び前記複数の記憶装置上への前記データバックアップ手段のバックアップデータの回復を行なうことを特徴とするデータ記憶システム。
【請求項１０】　請求項７において、
前記データバックアップ手段へのバックアップが実行されている前記複数の記憶装置上でのバックアップ対象ファイルシステムへのアクセスを記録しておく手段を有し、
上位コンピュータから該バックアップ対象ファイルシステムへのアクセスをしながら、該バックアップ対象ファイルシステムのデータのバックアップができることを特徴とするデータ記憶システム。
【請求項１１】　ディスクアレイコントローラと、ドラ

イブインターフェースによって該ディスクアレイコントローラに接続され、アレイ状に配列された複数のデータ等を格納するためのディスクドライブと、該ディスクドライブに格納されているデータをバックアップするためのデータバックアップ装置とを備え、
該複数のディスクドライブのうちの１以上のディスクドライブを予備ディスクドライブとし、残りをデータ等を格納するデータ／パリティディスクドライブとし、該予備ディスクドライブを、障害が生じた該データ／パリティディスクドライブに代ってデータ／パリティディスクドライブとするディスクアレイシステムにおいて、
データのバックアップの際、該予備ディスクドライブを該データ／パリティディスクドライブと該データバックアップ装置との間のデータ転送バッファとすることを特徴とするディスクアレイシステムのデータバックアップ方法。
【請求項１２】　請求項１１において、
前記データバックアップ装置は前記ディスクアレイコントローラによって制御されることを特徴とするディスクアレイシステムのデータバックアップ方法。
【請求項１３】　請求項１１において、
前記ディスクドライブはインターフェースを２ポート以上有し、前記ディスクアレイコントローラと前記データバックアップ装置とに接続されていることを特徴とするディスクアレイシステムのデータバックアップ方法。
【請求項１４】　請求項１１において、
前記アレイ上に配列された前記複数のディスクドライブ上での現在のファイルシステムの使用範囲を認識する手段もしくは上位コンピュータから、その認識結果を表わす情報を獲得し、
該上位コンピュータが介在することなく、該ファイルシステムの使用範囲のデータの前記データバックアップ装置へのバックアップの進行、及び前記アレイ上に配列された前記複数のディスクドライブ上への前記データバックアップ装置のバックアップデータの回復を行なうことを特徴とするディスクアレイシステムのデータバックアップ方法。
【請求項１５】　請求項１４において、
前回のバックアップ時から変更されたアドレスを認識する手段もしくは前記上位コンピュータから、その認識結果を表わす情報を獲得し、
前記上位コンピュータが介在することなく、前記データバックアップ装置への差分バックアップの進行、及び前記アレイ上に配列された前記複数のディスクドライブ上への前記データバックアップ装置のバックアップデータの回復を行なうことを特徴とするディスクアレイシステムのデータバックアップ方法。
【請求項１６】　請求項１４において、
前記データバックアップ装置へのバックアップが実行されている前記アレイ上に配列された前記複数のディスクドライブ上でのバックアップ対象ファイルシステムへのアクセスを記録しておき、
上位コンピュータから該バックアップ対象ファイルシステムへのアクセスをしながら、該バックアップ対象ファイルシステムのデータのバックアップができることを特徴とするディスクアレイシステムのデータバックアップ方法。
【請求項１７】　ディスクから読み出したデータを一時的に格納する半導体メモリを有するディスクドライバにおいて、
該ディスクの記録領域の所定部分を書込み禁止領域に設定する第１の手段と、
該半導体メモリの記録領域を複数に分割して管理する第２の手段とを備え、
該半導体メモリにおける該第２の手段によって分割形成された１乃至複数の記憶領域に、該第１の手段によって設定された書込み禁止領域の一部または全部から読み出したデータを記録して退避させ、しかる後に、該ディスクでの書込み禁止領域とした記録領域を書込み可能とすることを特徴とするディスクドライバ。
【請求項１８】　請求項１７において、
ホスト装置と接続する手段を複数設けたことを特徴とするディスクドライバ。
【請求項１９】　請求項１７または１８に記載のディスクドライバが複数個とアレイ制御装置とからなディスクアレイ装置において、
該ディスクドライバのデータバックアップの際、該アレイ制御装置の制御のもとに、データバックアップ領域に対する前記書込み禁止領域の設定と前記半導体メモリへのデータ退避動作とを行なうことを特徴とするディスクアレイ装置。
【請求項２０】　請求項１９において、
前記ディスクドライバと前記アレイ制御装置とを接続する経路を複数個設け、
前記書込み禁止領域の設定と退避動作との命令を伝達する前記接続経路とバックアップデータを前記ディスク装置の設定された書込み禁止領域から前記半導体メモリに転送する前記接続経路とを別にすることを特徴とするディスクアレイ装置。
【発明の詳細な説明】
【０００１】
【産業上の利用分野】本発明は、ディスクドライブ等のデータの記憶装置を所定個数備えたコンピュータシステムに係り、特に、該記憶装置に記憶されているデータのバックアップに関する。
【０００２】
【従来の技術】ディスクアレイ装置については、例えば特開平１－２５０１２８号公報に開示されている。この装置の特徴は、複数のディスクドライブが同時多重動作することによる高速化と、冗長ディスクドライブや予備

ディスクドライブを持つことによる高信頼化とを実現したことである。

【０００３】ディスクドライブを高速化するためには、データ転送速度の高速化及び１秒当りのリード／ライト処理件数（ＩＯスループット）の向上とを図ることが必要であるが、ディスクアレイ装置では、複数台のディスクドライブに対して並列にデータ転送を可能とすることにより、データ転送速度の高速化を実現している。かかる構成に、信頼度を上げるため、後述する冗長度を付加した構成を、一般に、ＲＡＩＤ３と呼んでいる。また、ＩＯスループットを向上させるために、複数台のディスクドライブに多重シーク／回転待ちさせるようにした構成があり、かかる構成に冗長度を付加して信頼度を上げるようにした構成を、ＲＡＩＤ５と呼んでいる。

【０００４】さて、以上のような性能を向上させるためには、上記いずれの構成においても、複数台のディスクドライブが必要となる。現在の標準的なディスクドライブの平均故障時間（ＭＴＢＦ）は５万時間程度であるが、このようにＮ台のディスクドライブを用いたディスクアレイ装置での平均故障時間（ＭＴＢＦ）は、５万／Ｎ時間に抑えられる。しかし、この値は、標準的な１０台構成のディスクアレイ装置の場合５千時間、即ち七ヶ月弱になり、システムとしては、一般に、許容し難い値である。

【０００５】この問題を解決するために、ディスクアレイ装置のディスクドライブを数台のデータディスクドライブと１台のパリティディスクドライブとからなるグループに区分し、また、データをストライプと呼ばれるデータ単位に分割し、上記グループ内の全ディスクドライブでの同じ論理アドレスから開始する全ストライプでパリティグループと呼ばれるグループを構成して、このパリティグループ内の全データストライプの排他的論理和を計算することによってパリティ情報を生成し、これをそのパリティグループのパリティディスクドライブにパリティストライプとして保存するようにした構成が知られている。かかる構成はＲＡＩＤと呼ばれるものであって、単にデータを複数のディスクドライブに分配するだけでなく、パリティグループ中の任意の１台が故障しても、パリティディスクドライブに保存されているパリティ情報から故障ディスクドライブのデータを回復することができるものであって、信頼性が向上する。

【０００６】ところで、ＲＡＩＤの平均故障時間（ＭＴＢＦ）は次の数１で表わされる。

【０００７】

【数１】

【数1】

$$MTBF = \frac{(MTBF)^2}{N \cdot G \cdot (G+1) \cdot MTTR}$$

MTBF … ディスクアレイ装置全体の平均故障時間

MTBF … ディスクアレイ装置に用いられるディスク単体の平均故障時間

MTTR … ディスクの平均修理時間

N … パリティグループの数

G … パリティグループ中のデータディスク数

【０００８】この数１によると、障害ディスクドライブを正常なディスクドライブと交換し、このディスクドライブへデータを回復するまでの時間、即ちディスクドライブの平均修理時間（ＭＴＴＲ）が平均故障時間（ＭＴＢＦ）を決定する重要なファクターになっていることが判る。この平均修理時間（ＭＴＴＲ）を最小限にするためには、障害ディスクドライブと交換する予備ディスクドライブを予めパリティグループ内に備えておき、障害が発生したとき、速やかに予備ディスクドライブを用いてデータの回復を行なうことが効果的である。かかる予備ディスクドライブについては、スペアディスクドライブとして、前記特開平１－２５０１２８号公報にも開示されている。

【０００９】ＲＡＩＤ構成の採用とスペアディスクドライブの採用とにより、複数台のディスクドライブからなるディスクアレイ装置での平均故障時間（ＭＴＢＦ）の低下の問題は解決される。

【００１０】

【発明が解決しようとする課題】しかし、上記数１で表わされる平均故障時間（ＭＴＢＦ）は、システム内の複数のディスクドライブの障害発生という事象が互いに独立であるという仮定をもとにしたものであり、地震・火災等の大規模災害の場合や、障害回復作業中のミスによる２重障害の発生の場合等は考慮されていない。また、オペレーションのミスによるデータの消失についても考慮されていない。

【００１１】かかる問題を解消するためには、データのバックアップを定期的に行なうこと以外に現実的方法はなく、上記従来技術では、ディスクドライブのアレイ化による容量の増大に対応した効果的なバックアップ方法については、配慮されていない。実際、標準的な５．２５”ディスクドライブ（１．５Ｇバイト）を使用し、かかるディスクドライブ１０台でパリティグループを構成してそのうちの１台をスペアディスクドライブに、他の１台をパリティディスクドライブに夫々に割り当てた場合、このパリティグループの総容量は１０ＧＢとなり、標準的なバックアップ用テープドライブ（例えば、１００ｋバイト／ｓｅｃあるいは約２００ｋバイト／ｓｅｃ）でバックアップするのに要する時間は１４時間弱である。これでは、ディスクアレイ装置の高ＩＯスループットを生かしたオンライントランザクションシステムに応用する場合、大きな障害となる。毎日データのバックアップを行なう場合、この１４時間、システムのサービスを停止してバックアップ作業を行わなければならないからである。

【００１２】本発明の目的は、かかる問題を解消し、データバックアップの占有時間を大幅に短縮して、かつデータバックアップを効果的に行なうことができるようにしたディスクドライブ，ディスクアレイ装置，データ記憶システム及びディスクアレイシステムのデータバックアップ方法を提供することにある。

【００１３】また、本発明の他の目的は、システムのサービスを閉塞せず、有効なデータバックアップを行なうことができるようにしたディスクドライブ，ディスクアレイ装置，データ記憶システム及びディスクアレイシステムのデータバックアップ方法を提供することにある。

【００１４】

【課題を解決するための手段】上記目的を達成するために、本発明は、データが記憶するための情報記憶装置に障害が発生したとき、この障害情報記憶装置に代えてデータを記憶するための情報記憶装置とする予備の記憶装置を、該データのバックアップ時、該情報記憶装置とデータバックアップ装置との間のデータ転送バッファとする。

【００１５】また、本発明は、前記データバックアップ装置でバックアップされている前記情報記憶装置上でのバックアップ対象ファイルシステムへのアクセス情報を記録しておく手段を有し、該アクセス情報をもとに、上位コンピュータから該バックアップ対象ファイルシステムへのアクセスをしながら、該バックアップ対象ファイルシステムのデータのバックアップができるようにする。

【００１６】さらに、本発明は、ディスクから読み出したデータを一時的に格納する半導体メモリを有するディスクドライバにおいて、該ディスクの記録領域の所定部分を書込み禁止領域に設定する第１の手段と、該半導体メモリの記録領域を複数に分割して管理する第２の手段とを備え、該半導体メモリにおける該第２の手段によって分割形成された１乃至複数の記憶領域に、該第１の手段によって設定された書込み禁止領域の一部または全部から読み出したデータを記録して退避させ、しかる後に、該ディスクでの書込み禁止領域とした記録領域を書込み可能とする。

【００１７】

【作用】コンピュータシステムにおけるアレイ上に配列された記憶装置には、ホストコンピュータが使用するデータ（ファイル）やこのデータから計算されるパリティ情報を格納するための情報記憶装置と、かかる情報記憶装置に障害が発生した場合にこの障害情報記憶装置と論理的に置き換えるための予備記憶装置とがある。本発明では、かかる予備記憶装置を、情報記憶装置でのデータをバックアップする際、情報記憶装置とデータバックアップ装置との間のデータ転送バッファとして用いる。

【００１８】これによると、データバックアップ装置のデータ転送速度が情報記憶装置のデータ転送速度よりも充分遅くても、情報記憶装置からデータ転送バッファへのデータ転送速度が速いため、少なくとも情報記憶装置からデータ転送バッファへのデータ転送は高速で行なわれ、従って、情報記憶装置に関しては、データバックアップに要する時間が非常に短くなる。

【００１９】現在大容量データバックアップ装置として一般に使用されている８ミリテープドライブやデータＤＡＴ（ディジタル・オーディオ・テープレコーダ）装置のデータ転送速度は２００ｋＢ／秒程度であり、ディスクドライブのデータ転送速度は３ＭＢ／秒程度であって、上記大容量データバックアップ装置と１０倍以上の差がある。このため、予備ディスクドライブでバックアップ装置へのデータ転送をバッファリングすることにより、ディスクドライブでのデータバックアップのための占有時間を１０分の１以下にすることができる。先に従来技術として挙げた例で説明すると、従来１４時間程度かかっていたデータバックアップのためのディスクドライブの占有時間を１．４時間以下にすることができるということである。これは、１日に１回データバックアップを行なうシステムでは、このバックアップのためのシステム停止時間が１日当り１４時間から１．４時間に減少するということである。

【００２０】また、本発明では、データバックアップ装置でバックアップされている情報記憶装置上でのバックアップ対象ファイルシステムへのアクセス情報を記録しておくことができるため、このバックアップ対象のファイルシステムを上位コンピュータ等によってアクセスされた後、このファイルシステムをデータバックアップ装置に転送することができる。従って、データバックアップ中でのアクセスがなされても、アクセス後のファイルシステムをバックアップすることができる。

【００２１】さらに、本発明では、ディスクの記録領域でのデータバックアップしようとする領域を書込み禁止領域とし、この書込み禁止領域から読み出したデータを一時的に半導体メモリに格納して退避させる。この書込み禁止領域から全てのデータが読み出されて半導体メモリに退避されると、この領域の書込み禁止が解除されて通常の外部からのデータ書込みが可能となる。半導体メモリに退避されたデータはバックアップ装置に転送される。そこで、ディスクのデータが読み出されてから最終的にバックアップ装置に転送されるまでに要する時間であるデータバックアップ転送時間は充分長くても、ディスクから半導体メモリへのデータ転送は高速で行なわれるから、書込み禁止が設定されても短時間でそれが解除され、従って、データバックアップの占有時間が短い。

【００２２】

【実施例】以下、本発明の一実施例を図面により説明する。図１は本発明によるディスクアレイ装置，データ記憶システム及びディスクアレイシステムのデータバックアップ方法の一実施例を示すブロック図であって、１はホストコンピュータ、１－１はＣＰＵ（中央処理ユニット）、１－２は主記憶部、１－３はＩ／Ｏバッファ、１－４はＩ／Ｏ（入出力）制御装置、１－５はディスクＩ／Ｆ（インターフェース）、１－６はＯＳ（オペレーティングシステム）、１－７はファイル管理部、１－８はデバイスドライバ、２はディスクアレイコントローラ、２－１はホストＩ／Ｆ、２－２はＣＰＵ、２－３はデータ転送制御装置、２－４はパリティ生成／チェック部、２－５は転送バッファ、２－６－１，２－６－２，２－６－３，２－６－４は夫々ドライブＩ／Ｆ、３－１，３－２，３－３は夫々バックアップ装置、４－１，４－２は夫々スペアディスクドライブ、５は データ／パリティディスクドライブ、６，７－１，７－２，７－３，７－４は夫々ＳＣＳＩ（Small Computer System Interface）バスである。

【００２３】同図において、ホストコンピュータ１、ディスクアレイコントローラ２及びバックアップ装置３－１がＳＣＳＩバス６で相互に接続されており、ディスクアレイコントローラ２には、ＳＣＳＩバス７－１，７－２，……，７－３により、ディスクドライブがアレイ状に接続されている。これらディスクドライブのうち、ＳＣＳＩバス７－１，７－２，……夫々に接続されるディスクドライブはデータ／パリティディスクドライブ５であって、これらのうちのＳＣＳＩバス７－１，７－２，……の１つに接続されたディスクドライブが上記のパリティディスクドライブであって、残りがデータディスクドライブである。また、ＳＣＳＩバス７－④に接続されているディスクドライブがスペアディスクドライブ４－１，４－２，……である。これらＳＣＳＩバス７－１，７－２，……，７－３夫々に接続されるディスクドライブの個数は等しく、各ＳＣＳＩバス７－１，７－２，…の同順位のディスクドライブ（即ち、図面上で横方向に１列に並ぶデータ／パリティディスクドライブ５）が上記のパリティグループを構成していて、各パリティグループにＳＣＳＩバス７－３に接続されたスペアディスクドライブ４が１つずつ付加されている。

【００２４】なお、バックアップ装置３－１の代りに、ＳＣＳＩバス７－１～７－３のいずれか、もしくはディスクアレイコントローラ２に別個のＳＣＳＩバス７－４を設けてこれに、バックアップ装置３－３または３－２を設けるようにしてもよい。

【００２５】以上のディスクアレイコントローラ２、アレイ状に配列されたディスクドライブ及びバックアップ装置によってディスクアレイ装置が構成されている。

【００２６】ホストコンピュータ１は、演算などの処理を行なうＣＰＵ１－１、ディスクドライブ５などへの入出力データをバッファリングやキャッシングするＩ／Ｏバッファ１－３を有し、ＣＰＵ１－１が使用するデータや命令を一時格納しておく主記憶部１－２、ディスクドライブ５などへのデータの入出力を制御するハードウェアであるＩ／Ｏ制御部１－４（最近のワークステーションでは、Ｉ／Ｏ制御部１－４に専用プロセッサを使用し、ディスクドライブなどへのデータの入出力処理のためのＣＰＵ１－１の負荷を軽減し、高速なＩ／Ｏ処理を実現するようにしている）、ディスクアレイコントローラ２とのインタフェース制御を行なうディスクＩ／Ｆ１－５及びＯＳ１－６からなっている。データ／パリティディスクドライブ５などの２次記憶装置へのデータの入出力処理は、ソフトウェア的には、ＯＳ１－６のファイル管理１－７、デバイスドライバ１－８などのモジュールが管理する。これらはアプリケーションプログラムが要求するデータ／パリティディスクドライブ５とのデータ入出力のスケジューリング、バッファリング、アクセスデータ長などを決定する。

【００２７】この実施例では、夫々のインターフェース（Ｉ／Ｆ）をＳＣＳＩとしている。ＳＣＳＩはＡＮＳＩ（American National Standard Institute）で作成され、ＩＳＯ（International Standard Organization）でも承認されたコンピュータ周辺機器用インタフェースであり、磁気ディスクドライブだけでなく、光ディスクドライブ、磁気テープ装置、プリンタ、スキャナなどもサポートしており、小型コンピュータに広く普及している。

【００２８】ディスクアレイコントローラ２は、そのドライブＩ／Ｆ２－６－１，２－６－２，……，２－６－３に接続されてアレイ状に配置されたデータ／パリティディスクドライブ５やスペアディスクドライブ４を統括制御し、ホストコンピュータ１に恰も通常のディスクドライブが１台接続しているかのように見せかける。ホストコンピュータ１から送られるデータはホストＩ／Ｆ２－１から取り込まれ、その論理アドレスに従い、データ

転送制御部２－３、転送バッファ２－５を介してアレイ状に配置されたデータ／パリティディスクドライブ５に分配される。ホストコンピュータ１からデータの要求があると、データ転送制御部２－３、転送バッファ２－５を介してホストＩ／Ｆ２－１から、アレイ状に配置されたデータ／パリティディスクドライブ５に分配して配置されているデータが統合されてホストコンピュータ１に転送される。

【００２９】また、ディスクアレイ装置では、同じパリティグループ内で１台のデータ／パリティディスクドライブ５が故障しても、そこでのデータの回復が行なえるようにするために、ホストコンピュータ１から送られてくるデータに冗長データ（即ち、パリティ情報）が付加され、データ／パリティディスクドライブ５のうちのパリティディスクドライブに格納する。このパリティ情報の生成とチェックはパリティ生成／チェック部２－４で行なわれる。データ／パリティディスクドライブが故障すると、そのパリティグループでのパリティディスクドライブに格納されているパリティ情報と故障していないデータディスクドライブのデータとで演算を行ない、故障したデータ／パリティディスクドライブのデータをこの計算によって生成して、このパリティグループに対するスペアディスクドライブ４に格納する。故障したデータディスクドライブでの全てのデータが回復してこのスペアディスクドライブ４に格納されると、このスペアディスクドライブは故障したデータディスクドライブの論理的位置を占め、この故障したデータディスクドライブに代ってデータディスクドライブとなる。

【００３０】このようにして、データディスクドライブの故障が発生しても、人手を介することなく、自動的に速やかにデータの回復処理を行なうことができる。なお、スペアディスクドライブは、データディスクドライブの故障が発生していない通常の場合、使用されない。

【００３１】バックアップ装置３－１，３－２または３－３は、ディスクアレイ装置に格納されているデータをバックアップするためのものである。上記のように、ディスクアレイ装置では、１パリティグループ内で任意の１台のデータディスクドライブが故障してもデータを失うことがないが、同じパリティグループ内で複数台のデータディスクドライブが同時に故障する多重障害が発生した場合や、オペレータミスによって誤ってデータが消去された場合にはデータを失ってしまう。バックアップ装置３－１，３－２または３－３はこのようなことを防止するために設けられたものであって、これにディスクアレイ装置に格納されているのと同じデータをバックアップしておき、上記のような障害が生じたとき、バックアップ装置のデータをディスクアレイ装置に送って（バックアップデータの回復）使用できるようにする。かかるバックアップ装置としては、磁気テープ装置や光磁気ディスクドライブなど記憶容量当りのコストが安い可換媒体の記憶装置が主に使用される。近年では、８ミリ磁気テープ装置で５ＧＢなど大容量の装置が開発されている。

【００３２】以下、この実施例の動作を説明する。

（１）ホストコンピュータ１がスペアディスクドライブとバックアップ装置とを認識する場合：図２はこの場合のホストコンピュータ１から見た場合の論理的な構成を示す図である。

【００３３】同図において、ここでの符号２は、ディスクアレイが図１でのディスクアレイコントローラ２によって統括、制御され、ホストコンピュータ１からは１台のディスクドライブに見せかけられているデータ／パリティディスクドライブ５のアレイの総体を概念的に表わしたものであり、これをディスクアレイということにする。また、バッファディスク８は図１におけるスペアディスク４（一般には、複数台）を概念的に表わしたものである。ホストコンピュータ１からは、これらがバックアップ装置３（図１でのバックアップ装置３－１，３－２，３－３のいずれか）とともに１つのＳＣＳＩバスＩＤを持った装置として、図２に示すように見える。なお、ＳＣＳＩバス ＩＤとは、ＳＣＳＩバス上に接続された複数台の装置を区別するための装置毎に割り振られたアドレスである。

【００３４】以上の構成の場合、ホストコンピュータ１は、バッファディスク８を１つのＳＣＳＩバス装置として認識しているため、バックアップに必要な転送を全て自分の管理の元で行なうことができる。但し、必ずしも、全て管理する必要がないことはいうまでもない。

【００３５】ディスクアレイ２のデータをバックアップ装置３にバックアップするのに行なうデータ転送は、ディスクアレイ２からバッファディスク８への転送（１）とバッファディスク８からバックアップ装置３への転送（２）とからなっている。好適な例としては、転送（１）はホストコンピュータ１がディスクアレイ２のファイル構造を認識しながら行ない、転送（２）は起動指示と終了確認のみホストコンピュータ１が介在する方法である。これによると、転送（２）中では、ホストコンピュータ１はディスクアレイ２に対して通常のアクセスを行なうことができる。

【００３６】例えば、ＵＮＩＸマシンの場合には、ＴＡＲコマンドを用いてバッファディスク８への転送を行なう。この際、バッファディスク８はＲＡＷデバイスとしてアクセスされる。ＴＡＲコマンドで転送した場合、通常のＭＴ装置へバックアップする場合と同じフォーマットになるため、データ転送（２）では、ホストコンピュータ１からバッファディスク８に対して転送すべき範囲と転送先さえ起動時に指示すればよい。但し、そのためには、バッファディスク８がＳＣＳＩバスコマンドの１つであるＣＯＰＹコマンドをサポートしていることが必要である。

【００３７】上記転送（１）において、その開始から終了までホストコンピュータ１が介在する理由は、ａ）オープンしているファイルの処理、ｂ）バックアップすべきデータが存在する領域の判断、ｃ）ディスク上のデータフォーマットと通常のバックアップ装置（磁気テープ装置など。通常、これらはシーケンシャルアクセスデバイスであって、ランダムアクセスデバイスである磁気ディスクとは異なるフォーマットを持っている）のフォーマット変換などの処理を通常サポートしているコマンドのみで、上記転送（１）を行なうことができるためである。転送（１）をホストコンピュータ１の介在なしで行なう場合には、後に述べるように、上記ａ），ｂ），ｃ）の処理，判断ためのコマンド，メッセージなどをサポートする必要がある。

【００３８】但し、図１に示すように、ディスクアレイコントローラ２は、実際には１つのＳＣＳＩバス装置であるのに、ホストコンピュータ１に対しては、図２におけるディスクアレイ２とバッファディスク８の二役をしなければならない。ＳＣＳＩバス規約にはこのような構成は存在しないが、実施は可能である。

【００３９】（２）ホストコンピュータ１は、バックアップ装置を認識するが、スペアディスク装置を認識しない場合：図３はこの場合のホストコンピュータから見た論理的構成図であり、図２に対応する部分には同一符号を付けている。

【００４０】ここでは、図２でのバッファディスク８はホストコンピュータ１によって認識されていない。この場合、上記のデータ転送（１）に当たるデータ／パリティディスクドライブ５からスペアディスクドライブ４へのデータ転送は、ディスクアレイコントローラ２が起動、終了処理を含めて全て管理する。この場合、通常のバックアップ用のコマンド（上記の例では、ＴＡＲコマンド）を用いてバックアップを行なうことも考えられる。それでも、一旦スペアディスクドライブ４でバッファリングすることにより、通常のアクセスのためにデータ／パリティディスクドライブ５を使用できる時間が増大する効果がある。

【００４１】また、バックアップのために必要なファイル管理情報を読み込むためには、ディスクアレイ２でのデータをホストコンピュータ１が読み込むが、バックアップ装置３にデータを転送するためには、ホストコンピュータ１にデータを読み込まず、ＳＣＳＩバスのＣＯＰＹコマンドを用いてディスクアレイ２のデータをバックアップする方法がある。ＣＯＰＹコマンドを受け取ったときには、スペアディスクドライブ４をデータ転送バッファとして用い、ディスクアレイコントローラ２が制御を行なうことにより、より高速なバックアップを行なうことができる。

【００４２】ディスクアレイコントローラ２でのバックアップ処理は、バックアップ装置３の物理的接続位置により、図４、図５及び図６に示す３種類の好適な例を挙げることができる。

【００４３】図４はバックアップ装置３をホストＩ／Ｆ２－１側のＳＣＳＩバス６に接続した場合であり、図５は他のディスクドライブが接続されていないドライブＩ／Ｆ２－６に接続した場合であり、図６はスペアディスクドライブ４が接続されているドライブＩ／Ｆ２－６に接続した場合である。なお、図４～図６においても、データ／パリティディスクドライブ５及びスペアディスクドライブ４が夫々１つずつ示されているが、勿論これらは図１で説明したように複数台アレイ状に配置されており、ディスクアレイコントローラ２によって１台のディスクドライブのように統括制御されているものを概念的に表わしているものである。

【００４４】図４に示す例の場合、ホストコンピュータ１はバックアップを行なうデータ／パリティディスクドライブ５の領域のファイル管理情報を読み込み、データ／パリティディスクドライブ５からバックアップ装置３への転送を指示するコマンドをディスクアレイコントローラ２に対して発行する（その転送はホストコンピュータ１が介在してもよい）。ディスクアレイコントローラ２は転送先がバックアップ装置３であるデータを一旦スペアディスクドライブ４に転送し、次に、このスペアディスクドライブ４からバックアップ装置３への転送を行なう。これにより、データ／パリティディスクドライブ５がバックアップのために占有される時間を減らすことができる。

【００４５】図５に示す例の場合には、バックアップ装置３がディスクアレイコントローラ２のドライブＩ／Ｆ２－６に接続されているため、スペアディスクドライブ４からバックアップ装置３へのデータ転送はホストＩ／Ｆ２－１側のＳＣＳＩバス６を通らずに行なわれる。このため、バックアップ中のホストコンピュータ１からデータ／パリティディスクドライブ５へのアクセス性能の劣化が小さい。

【００４６】但し、この場合、ホストコンピュータ１に対して図３に示したようにバックアップ装置３が接続されていると見せるためには、ディスクアレイコントローラ２は、バックアップ装置３に対するホストコンピュータ１からのアクセスを透過的に受渡してやらなければならない。このことは、図２におけるバッファディスク８と同様、ＳＣＳＩバスの規約には規定されていない方法ではあるが、実施は可能である。

【００４７】図６に示す例は、バックアップ装置３をスペアディスクドライブ４と同じＳＣＳＩバス上に接続する場合である。この場合には、スペアディスクドライブ４からバックアップ装置３へのデータ転送がディスクアレイコントローラ２の内部も通過しないで行なわれ、このために、さらにバックアップ中のホストコンピュータ１からデータ／パリティディスクドライブ５へのアクセ

ス性能劣化を防止できる。かかる効果は、スペアディスクドライブ４がＣＯＰＹコマンドをサポートすることによって最大になる。

【００４８】（３）ホストコンピュータ１がバックアップ装置もスペアディスクも認識しない場合：図７はこの場合のホストコンピュータ１からみた概念的な構成図である。実際の物理的な接続は、バックアップ装置３を認識する場合と同様に図４，図５及び図６に示す３種類が好適な実施例として挙げられる。この場合、ホストコンピュータ１がバックアップ装置３を認識しないため、ディスクアレイコントローラ２はバックアップ用のコマンドをサポートしなければならない。なお、かかるコマンドはＳＣＳＩバスで規定されていないが、このようなコマンドでも、ベンダユニークとして実施することができる。バックアップを行なうためには、まず、ホストコンピュータ１がバックアップ対象領域のファイル管理情報を読み込み、バックアップすべきアドレスとバックアップを指示するコマンドをバックアップを行なう順にディスクアレイ２に発行してやればよい。

【００４９】（４）２ポートＳＣＳＩバスディスクドライブを使用した場合：以上説明した各実施例では、ディスクドライブ装置として通常の１ポートのものを想定していた。しかし、以上全ての実施例において、２ポートＳＣＳＩバスディスクドライブを使用した構成も考えられる。その一実施例を図８に示す。かかる実施例の特徴は、データ／パリティディスクドライブ５からスペアディスクドライブ４へのデータ転送やコマンド／メッセージの転送が、通常のホストコンピュータ１からデータ／パリティディスクドライブ５へのアクセスとＳＣＳＩバス上で競合しない点である。スペアディスクドライブ４からバックアップ装置へのデータ転送も、バックアップ装置３が点線で示されている位置に接続されている場合には、ホストコンピュータ１からデータ／パリティディスクドライブ５へのアクセスと競合することがなく、バックアップ中の性能劣化が少ない。

【００５０】（５）システムを閉塞せずにバックアップを行なう方法：以上説明した実施例では、バックアップのためにディスクドライブを占有する時間を短縮することができた。バックアップを行なっているファイルシステムもしくはパーティションは、バックアップ中データ内部の統一性や管理情報との統一性が失われないようにするために、少なくとも一纏まりのファイルとその管理情報の処理中、その領域へのライトアクセスを停止（システムの閉塞）しなければならない。しかし、オンライントランザクションの用途では２４時間のサービスが要求される場合もあり、このようなシステムでは、上記のような閉塞は許容できない。

【００５１】このようなシステムの閉塞を行なわずにバックアップを行なうためには、以下の手順に従えばよい。即ち、１）バックアップ中にバックアップ対象のディスク領域の中でオープンされているファイルを記録する。記録先はディスク上でも、メモリ上でもよい。オープンしているファイルのリストは、一般に、ホストコンピュータの主記憶部上に存在する。２）対象領域をオープンされているファイルも含めて（勿論、省いてもよい）、バックアップ装置にバックアップする。３）ファイルがクローズされたときバックアップ対象領域であったファイルならば、差分バックアップの要領で、２）で行なったバックアップに連続した領域にバックアップを行なう。

【００５２】なお、１）の代りに、ファイルオープン処理以降の全てのライトアクセスを記録し、ライトアクセスのない領域は、オープンされていても、バックアップを行ない、ライトアクセスのあった領域のみ、クローズ処理後、差分バックアップを行なうようにしてもよい。

【００５３】図９は本発明によるディスクドライブの一実施例を示すブロック図であって、８ａ，８ｂはＳＣＳＩバス、９ａ，９ｂはＳＣＳＩ制御部、１０はメモリ制御部、１１はデータメモリ、１２はフォーマット制御部、１３はリード／ライト用アンプ（以下、Ｒ／Ｗアンプという）、１４は磁気ヘッド、１５は記録ディスク、１６は書込禁止領域判定部、１７はスピンドルモータ、１８はＶＣＭ、１９はバッテリである。

【００５４】図９において、ＳＣＳＩ制御部９ａはＳＣＳＩバス８ａのプロトコル制御を行ない、ＳＣＳＩ制御部９ｂはＳＣＳＩバス８ｂのプロトコル制御を行なう。メモリ制御部１０は、ＳＣＳＩ制御部９ａ，９ｂとデータメモリ１１とフォーマット制御部１２との間のデータ転送を制御し、図１０に示すように、データメモリ１１を４つの記憶領域に分割して管理する。データメモリ１１は少なくともディスク１５の数トラック分の容量を有していることが望ましく、ここでは、その容量は１Ｍバイトとする。データメモリ１１は、バッテリ１９により、データの不揮発化を行なっている。フォーマット制御部１２は、ディスク１５上の記録トラックのセクタ分割方法を管理する。Ｒ／Ｗアンプ１３は、ディスク１５へのデータ記録時、データによって磁気ヘッド１４を駆動し、ディスク１５からのデータ再生時、磁気ヘッド１４からの再生データ信号を増幅してディジタル信号に変換する。書込禁止領域判定部１６は、ディスク１５上のデータ書込み予定領域が書込み禁止領域か否かを判定し、書込み禁止領域ならば、フォーマット制御部１２の書込み動作を停止させる。

【００５５】次に、ディスク１５上の記録領域の所定領域のデータをバックアップする場合のこの実施例の動作を図１１を用いて説明する。

【００５６】ここで、データバックアップ動作は、バックアップの領域をデータメモリ１１へ退避させるためのロック命令と、データメモリ１１に退避したデータを外部のバックアップ装置に転送させるためのバックアップ

リード命令との２種類の命令をディスクドライブが図示しないホスト装置から受け取ることにより、実行される。

【００５７】ディスクドライブは、ホスト装置から部分ロック命令を受けると（ステップ１０１）、ディスク１５のこの部分ロック命令で指定されるロック領域を一時的に書込み禁止領域に設定し（ステップ１０２）、かつこのロック領域中のデータを一時的に退避する領域をデータメモリ１１内に確保する（１０３）。データメモリ１１は、通常、図１０の「通常時」として示す状態にあるが、部分ロック命令があると、ロック領域の容量がデータメモリ１１の容量の３／４以下ならば、ロック領域のデータを格納できる最小領域を退避領域として、図１０に示す「退避時１」，「退避時２」，「退避時３」のいずれかの状態に設定される。ロック領域の容量がデータメモリ１１の容量の３／４を越えるならば、データメモリ１１を図１０の「退避時３」の状態とし、データメモリ１１の容量の３／４に等しい容量のデータをロック領域から読み出し、データメモリ１１の第２～第４領域の退避領域に格納し（ステップ１０４）、読出しが終了したディスク１５のロック領域の書込みを許可する（ステップ１０５）。データメモリ１１の退避領域にロック領域の全データを格納できない場合には（ステップ１０６）、部分ロックの完了をＳＣＳＩ制御部１０に通知する（ステップ１０７）。また、データメモリ１１の退避領域にロック領域の全データを格納できた場合には（ステップ１０６）、ロック対象領域の退避が完了すると、書込み可能であることを通知する（ステップ１０８）。

【００５８】その後、バックアップリード命令を受けると（ステップ１０９）、ディスクドライブは、このバックアップリード命令が転送されてきたＳＣＳＩバスがＳＣＳＩバス８ａ，８ｂのいずれかを介し、データメモリ１１の退避領域からバックアップデータをホスト装置へ転送する（ステップ１１０）。ホスト装置からバックアップデータの受信完了の通知を受けると（ステップ１１１）、ディスクドライブは、前記の部分ロック命令で指定されたバックアップ領域のバックアップデータの全てをホスト装置へ転送終了したならば（ステップ１１２）、データメモリ１１に設定された退避用領域を通常用領域に戻し（ステップ１１３）、ディスク１５での部分ロック状態を解除し（ステップ１１４）、バックアップ動作を完了する。

【００５９】バックアップ領域に未転送部分が残っているときには（ステップ１１２）、引き続きステップ１０４～１１２の一連の動作を行ない、残りのロック領域のデータに対してデータメモリ１１への退避（ステップ１０４）とホスト装置への転送（ステップ１１０）を繰り返す。

【００６０】なお、ステップ１０７，１０８を経た後、バックアップリード命令がないときに（ステップ１０９），ディスク１５への読出し命令を受けると（ステップ１２０），通常の読出し処理が行なわれ、ディスク１５への書込み命令を受けると（ステップ１２２），書込み禁止領域であるときには（ステップ１２４），書込みを禁止するが（ステップ１２６）、書込み禁止領域でないときには通常の書込み処理を行なう（ステップ１２５）。読出し、書込み命令のいずれをも受けないときには、他の命令の処理を行なう（ステップ１２３）。

【００６１】以上のようにして、この実施例では、バックアップデータをデータメモリ１１に退避した後には、ディスク１５へのデータの書込み動作とバックアップのためのディスクドライブのスループットの低下を抑えることができる。

【００６２】図１２は図９～図１１で説明したディスクドライブをアレイ上に配列してなる本発明によるディスクアレイ装置のさらに他の実施例を示すブロック図であって、２０はホストコンピュータ、２１はディスクアレイ装置、２２はホストインタフェイス制御部（以下、ホストＩ／Ｆという）、２３はデータ分配制御部、２４はキャッシュメモリとしてのバッファメモリ、２５はパリティ生成部、２６はデータ回復部、２７は磁気テープ装置制御部（以下、ＭＴ制御部という）、２８は磁気テープ装置（以下、ＭＴ装置という）、２９ａ～２９ｅはＳＣＳＩ制御部、３０ａ～３１ｅ，３２は図９で説明したディスクドライブ（以下、ＨＤＤという）、３３，３４はＳＣＳＩ制御部である。

【００６３】図１２において、ホストＩ／Ｆ２２は、ホストコンピュータ２０からディスクアレイ装置２１への命令や書込みデータの受信と、ディスクアレイ装置２１からホストコンピュータ２０への命令実行結果や読出しデータの送信を行なう。データ分配制御部２３は、ホストＩ／Ｆ２２とキャッシュメモリ２４とパリティ生成部２５とデータ回復部２６とＭＴ制御部２７とＳＣＳＩ制御部２９ａ～２９ｅ，３３，３４との間のデータ転送の制御を行なう。さらに、このデータ分配制御部２３は、データ書込み時には、ホストコンピュータ２０からの書込みデータを４ｋバイト単位のデータブロック（先のデータストライプ）に分割して、ＨＤＤ３１ａ～３１ｅへ振り分け、逆に、データ読出し時には、ＨＤＤ３０ａ～３１ｅからのデータブロックをまとめて、ホストコンピュータ２０への読出しデータに再構成する。

【００６４】各データブロックは次のように書き込まれる。即ち、いま、データが４個のデータブロックＤ００，Ｄ０１，Ｄ０２，Ｄ０３に分割されたとすると、図１３に示すように、データブロックＤ００がＨＤＤ３０ａに、データブロックＤ０１がＨＤＤ３０ｂに、データブロックＤ０２がＨＤ３０ｃに、データブロックＤ０３がＨＤＤ３０ｄに順に配置されることになり、ＨＤＤ３０ｅには、これらデータブロックＤ００，Ｄ０１，Ｄ０２，Ｄ０３から生成したパリティブロックＰ０が配置さ

れる。次のデータがデータブロックＤ１０，Ｄ１１，Ｄ１２，Ｄ１３に分割され、それらのパリティブロックをＰ１とすると、これらも同様にＨＤＤ３０ａ～３０ｅに記憶されるが、パリティブロックＰ１はＨＤＤ３０ｄに記憶される。以下同様にして、順次のデータが記憶される。なお、同一データから分割されたデータブロックとこれらによって生成されたパリティブロックとをまとめてパリティグループということは、先に述べたとおりである。

【００６５】パリティブロックは、パリティ生成部１３が、次の数２に示すように、同一パリティグループの４つのデータブロックを排他論理和することにより、生成される。

【００６６】

【数２】

【数2】

$$P0 = D00 \oplus D01 \oplus D02 \oplus D03$$

【００６７】キャッシュメモリ１２は書込みデータや読出しデータを一時的に格納するものである。また、データ回復部１４は、ＨＤＤの故障のためにパリティグループ内のデータブロックが損失した場合、パリティグループ内の正常なデータブロックとパリティブロックを用いて、損失したデータブロックを回復するためのものである。例えば、ＨＤＤ３０ａが故障した場合、次の数３に示すように、データブロックＤ０１，Ｄ０２，Ｄ０３とパリティブロックＰ０とを用いて、データブロックＤ００を回復できる。

【００６８】

【数３】

【数3】

$$D00 = D01 \oplus D02 \oplus D03 \oplus P0$$

【００６９】ＳＣＳＩ制御部２９ａ～２９ｅ，３３，３４はＨＤＤを接続しているＳＣＳＩバスのプロトコル制御を行なう。ＭＴ制御部２７はバックアップ装置としてのＭＴ装置２８を制御し、ＨＤＤからのデータをバックアップのため、ＭＴ装置１６内の磁気テープに記録する。ＨＤＤ３０ａ～３１ｅ，３２は２系統のＳＣＳＩバスに接続可能である。また、ＨＤＤ３３はスペア用である。

【００７０】以下、この実施例のデータ記録再生動作を図１４を用いて説明するが、まず、ホストコンピュータ２０からのデータをＨＤＤ３０ａのＤ００領域とＨＤＤ３０ｂのＤ０１領域とに記録する場合を説明する。

【００７１】時刻ｔ０１にホストコンピュータ２０から書込み命令と書込みデータが送信されると、ホストＩ／Ｆ２２はこれら書込み命令と書込みデータをデータ分配制御部２３へ転送する。時刻ｔ０２で、データ分配制御部２３は書込みデータをキャッシュメモリ２４へ転送し、また、パリティブロックの生成のために、ＳＣＳＩ制御部２９ａ，２９ｂ及び２９ｅへデータ読出しを要求する。そこで、ＳＣＳＩ制御部２９ａはＨＤＤ３０ａへデータブロックＤ００の読取りを要求し、ＳＣＳＩ制御部２９ｂはＨＤＤ３０ｂへデータブロックＤ０１の読出しを要求し、ＳＣＳＩ制御部２９ｅはＨＤＤ３０ｅへパリティブロックＰ０の読出しを要求する。

【００７２】時刻ｔ０３で、ＨＤＤ３０ａ，３０ｂ，３０ｅは、夫々、同一パリティグループのデータブロック，パリティブロックを読み取り、ＳＣＳＩ制御部２９ａ，２９ｂ，２９ｅへこれらデータブロックを転送する。そこで、ＳＣＳＩ制御部２９ａ，２９ｂ，２９ｅは、これら読取りデータブロックをデータ分配制御部２３へ転送し、データ分配制御部２３はこれらをキャッシュメモリ１２へ格納する。

【００７３】ここで、ＨＤＤ３０ａ，３０ｂ，３０ｅからの読取りデータブロックとパリティブロックを、夫々、ｒｄ０，ｒｄ１，ｒｐ０とし、ホストコンピュータ２０からの書込みデータのデータブロックをｗｄ０，ｗｄ１とする。

【００７４】時刻ｔ０４で、キャッシュメモリ２４から、データブロックｒｄ０，ｒｄ１，パリティブロックｒｐ０，データブロックｗｄ０，ｗｄ１がパリティ生成部２５へ転送される。パリティ生成部２５では、次の数４のように、これらから新しいパリティブロックｗｐ０が生成され、キャッシュメモリ２４へ転送される。

【００７５】

【数４】

【数4】

$$wp0 = rd0 \oplus rd1 \oplus rp0 \oplus wd0 \oplus wd1$$

【００７６】時刻ｔ０５で、データ分配制御部２３は、ＳＣＳＩ制御部２９ａ，２９ｂ，２９ｅへ、書込み命令と、夫々、書込みデータブロックｗｄ０，ｗｄ１とパリティブロックｗｐ０を送る。これにより、ＳＣＳＩ制御部２９ａはＨＤＤ３０ａへデータブロックｗｄ０の記録領域Ｄ００への書込みを要求し、ＳＣＳＩ制御部２９ｂはＨＤＤ３０ｂへデータブロックｗｄ１の記録領域Ｄ０１への書込みを要求し、ＳＣＳＩ制御部２９ｅはＨＤＤ３０ｅへパリティブロックｗｐ０の記録領域Ｐ０への書込みを要求する。

【００７７】時刻ｔ０６で、ＨＤＤ３０ａ，３０ｂ，３０ｅのライト動作の完了がＳＣＳＩ制御部２９ａ，２９ｂ，２９ｅを通じてデータ分配制御部２３へ通知され、これにより、データ分配制御部２３は、ホストＩ／Ｆ２２を経由して、ホストコンピュータ２０へ書込み命令の完了を通知する。以上により、データ記録動作は終了する。

【００７８】次に、例えば、ＳＣＳＩ制御部２９ｂまたはそれに接続されるＳＣＳＩバスが故障した場合の、Ｈ

ＤＤ３０ａのＤ００領域とＨＤＤ３０ｂのＤ０１領域とからのデータの再生動作を説明する。

【００７９】いま、時刻ｔ１０でかかる故障が生じたとすると、ＳＣＳＩ制御部２９ｂはこのことをデータ分配制御部２３へ通知する。その後、時刻ｔ１１で、ホストコンピュータ２０から読取り命令が送信されるたとすると、ホストＩ／Ｆ２２は読取り命令をデータ分配制御部２３へ転送する。時刻ｔ１２で、データ分配制御部２３はＳＣＳＩ制御部２９ａへＨＤＤ３０ａの読取りを要求する。さらに、データ分配制御部２３は、故障中のＳＣＳＩ制御部２９ｂの代わりに、ＳＣＳＩ制御部３３へＨＤＤ３０ｂの読取りを要求する。

【００８０】時刻ｔ１３で、ＨＤＤ３０ａ，３０ｂが、夫々、ＳＣＳＩ制御部２９ａ，３３を介して、データ分配制御部２３へ読取りデータブロックを転送する。そこで、データ分配制御部２３はＨＤＤ３０ａ，３０ｂからのデータブロックを供給し、ホストコンピュータ２０へ送信する。これにより、データ再生動作が終了する。

【００８１】次に、この実施例のデータ回復動作を図１５を用いて説明する。時刻ｔ２０でＨＤＤ３０ａが故障したとすると、ＨＤＤ３０ａはこれを検出し、ＳＣＳＩ制御部２９ａを介して、データ分配制御部２３へこの旨通知する。時刻ｔ２１で、データ分配制御部２３は、データ回復部２６へＨＤＤ３０ａ内のデータブロックのスペアＨＤＤ３２への回復を要求する。データ回復部２６は、時刻ｔ２２で、ＳＣＳＩ制御部３３を介し、ＨＤＤ３０ｂ～３０ｅへ、先頭のパリティグループのデータブロックＤ０１，Ｄ０２，Ｄ０３及びパリティブロックＰ０の読取りを要求する。

【００８２】時刻ｔ２３から、データ回復部２６は、ＨＤＤ３０ａ～３０ｅからの読取りデータを受け、順次上記数２の演算を実行し、データブロックＤ００を回復させる。そして、時刻ｔ２４で、先頭のパリティグループの各ブロックＤ０１，Ｄ０２，Ｄ０３，Ｐ０の転送とデータ回復演算が完了すると、データ回復部２６は、ＳＣＳＩ制御部３３を介し、スペアＨＤＤ３２へ回復したデータブロックＤ００を書き込む。時刻ｔ２４で、スペアＨＤＤ３２での書込みが完了すると、この旨をＳＣＳＩ制御部３３を介してデータ回復部２６へ通知する。データ回復部２６は、データ分配制御部２３へデータブロックＤ００の回復完了を通知し、データブロックＤ００のアクセスを許可する。引き続き他のデータブロックについて、順次回復処理を行なう。

【００８３】時刻ｔ２６で、データ回復部２６は、最後のデータブロックＤｋ０の回復演算を完了し、ＳＣＳＩ制御部３３を介してスペアＨＤＤ３２へ回復したデータブロックＤｋ０を書き込む。そして、時刻ｔ２７で、スペアＨＤＤ３２が書込み完了をＳＣＳＩ制御部３３を介してデータ回復部２６へ通知する。データ回復部２６は、データ分配制御部２３へ全データブロックが回復完了したことを通知し、データ回復処理が完了する。

【００８４】以上のデータ回復処理においては、ＳＣＳＩ制御部２９ａ～２９ｅが使用されないため、データ回復処理中にも、故障したＨＤＤ３０ａのパリティグループ以外のＨＤＤ３１ａ～３１ｅには並行してデータブロックの書込み／読取りが可能であり、ＨＤＤ３０ｂ～３０ｅには読取りが可能である。

【００８５】次に、この実施例のデータバックアップ動作を図１６を用いて説明する。時刻ｔ３０で、ホストコンピュータ２０からディスクアレイ装置２１内のデータブロックＤ１２～Ｄ５３の領域の部分バックアップ命令があったとする。そこで、データ分配制御部２３は、ＳＣＳＩ制御部２９ａ～２９ｅを介して、ＨＤＤ３０ａ～３０ｅへデータブロックＤ１２～Ｄ５３の部分ロックを要求し、ホストコンピュータ２０からのかかる領域への書込み要求の受付を禁止する。各ＨＤＤは、前記のように１Ｍバイトのデータメモリ１１（図９）を持つため、このデータメモリ１１の退避領域にロック領域のデータを全て格納して退避させることができ、このロック対象領域の各ブロックの退避が完了すると、かかる領域での書込みが可能であることを通知する。この実施例では、１つのパリティグループで最大３８４０ｋバイトのデータを退避できる。

【００８６】次に、時刻ｔ３２で、ＳＣＳＩ制御部２０ａ～２０ｅの全てがＨＤＤの部分ロックの完了を、データ分配制御部２３へ通知し、これにより、データ分配制御部２３はＳＣＳＩ制御部３３にＨＤＤ３０ａ～３０ｅからＭＴ装置２８へのバックアップデータの転送を要求する。さらに、ホストコンピュータ２０から指定されたバックアップ領域が全てＨＤＤ３０ａ～３０ｅ内のデータメモリ１１内に格納できた場合には、ホストコンピュータ２０からのこの領域への書込み要求の受付を許可する。

【００８７】ＳＣＳＩ制御部３３は、時刻ｔ３３に、部分バックアップの先頭データブロックＤ１２の読取りをＨＤＤ３０ｃへ要求する。データブロックＤ１２を受けると、ＳＣＳＩ制御部３３は、時刻ｔ３４に、ＭＴ制御部２７へデータブロックＤ１２を送り、ＭＴ装置２８へ書込みを要求する。これにより、ＭＴ制御部２７はデータブロックＤ１２のＭＴ装置２８への書込みを開始する。

【００８８】ＳＣＳＩ制御部３３は、データブロックＤ１２の転送が終了すると、時刻ｔ３５に、ＨＤＤ３０ｄへデータブロックＤ１３の読取りを要求し、データブロックＤ１２のＭＴ装置２８への書込み完了の前に、予め、次のデータブロックＤ１３をＳＣＳＩ制御部３３へ読み出しておく。時刻ｔ３６に、ＭＴ制御部２７からデータブロックＤ１２の書込み完了通知があると、ＳＣＳＩ制御部３３はＨＤＤ３０ｃへデータブロックＤ１２の受信完了を通知する。

【００８９】続いて、ＳＣＳＩ制御部３３は、読み取った上記のデータブロックＤ１３のMT装置２８への書込み要求と、データブロックＤ２０のHDD３０ａからの読取り要求とを出す。時刻ｔ３７に、MT制御部２７からデータブロックＤ１３の書込み完了通知を受けると、ＳＣＳＩ制御部３３は、HDD３０ｄへデータブロックＤ１３の受信完了を通知し、データ分配制御部２３に第１列のパリティグループのデータブロックのバックアップ終了を通知する。

【００９０】同様に、時刻ｔ３８に、MT制御部２７からデータブロックＤ２３のライト完了通知を受けると、ＳＣＳＩ制御部３３は、データ分配制御部２３に第２列のパリティグループのデータブロックのバックアップ終了を通知する。時刻ｔ３９に、MT制御部２７からバックアップ領域の末尾のデータブロックＤ５３のライト完了通知を受けると、ＳＣＳＩ制御部３３は、データ分配制御部２３に第２列のパリティグループのデータブロックのバックアップ終了を通知し、時刻ｔ４０に、ホストコンピュータ２０に部分バックアップの完了を通知する。

【００９１】以上、バックアップ領域がHDD内のデータメモリ１１（図９）に全て退避できる場合を説明したが、バックアップ領域の方が大きい場合には、データ分配制御部２３は、ディスク１５（図９）のバックアップ領域のうち、データメモリ１１に退避された領域からホストコンピュータ２０からの書込み要求の受付を許可する。

【００９２】以上のように、バックアップデータの転送に際しては、ＳＣＳＩ制御部２９ａ～２９ｅは使用されないため、バックアップ処理中にも、並行して、HDD３０ａ～３１ｅへのアクセスが可能である。

【００９３】この実施例では、通常のHDDアクセス用バスに加えて、データ回復及びデータバックアップ用に使用するバスを持っているため、データ回復またはデータバックアップのためのバスの占有による性能低下を防ぐことができる。さらに、データ回復及びデータバックアップ用に使用するバスはスペア用HDDにも使用するため、スペアHDD３２の制御のために別にＳＣＳＩ制御部を設ける必要がなく、この分コストを削減できる。

【００９４】また、この実施例では、１つのパリティグループを１本のバスで接続しているため、HDDを順次読み出すだけで、パリティグループ内の複数のHDDに分散して配置されたデータブロックを元の順序に並べ変えることができる。

【００９５】

【発明の効果】以上説明したように、本発明によれば、バックアップ中のデータディスクの占有時間をバックアップ装置と磁気ディスクドライブの転送速度の比だけ減らすことができる（典型的な例では、磁気テープ装置が約２００ｋＢ／ｓｅｃであるのに対し、磁気ディスクドライブでは３MB／ｓｅｃと１／１５程度である）。

【００９６】また、本発明によれば、バックアップ対象領域のオープンされているファイルを記録し、クローズ後に再度バックアップを行なうことにより、システムを閉塞せずにバックアップを行なうことができる。

【００９７】さらに、本発明によれば、ディスクドライブ内の半導体メモリの一部にバックアップデータを退避しておき、通常のディスクアクセスの合間に、バックアップ用の装置に転送するので、データバックアップ時にディスクドライブが長時間占有されることを防ぐことができる。また、ディスクのバックアップ領域のうちデータ退避の終わっていない領域は書込み禁止となっているため、そこに書き込まれているデータを誤って壊すこともない。

【図面の簡単な説明】

【図１】本発明によるディスクアレイ装置，データ記憶システム及びディスクアレイシステムのデータバックアップ方法の一実施例を示すブロック図である。

【図２】図１に示した実施例でのホストコンピュータ側からみた論理的接続の一具体例を示す構成図である。

【図３】図１に示した実施例でのホストコンピュータ側からみた論理的接続の他の具体例を示す構成図である。

【図４】図３に示した論理的接続でのディスクアレイコントローラ側の処理の一具体例を示す図である。

【図５】図３に示した論理的接続でのディスクアレイコントローラ側の処理の他の具体例を示す図である。

【図６】図３に示した論理的接続でのディスクアレイコントローラ側の処理のさらに他の具体例を示す図である。

【図７】図１に示した実施例でのホストコンピュータ側からみた論理的接続のさらに他の具体例を示す構成図である。

【図８】本発明によるディスクアレイ装置，データ記憶システム及びディスクアレイシステムのデータバックアップ方法の他の実施例を示すブロック図である。

【図９】本発明によるディスクドライブの一実施例を示すブロック図である。

【図１０】図９でのデータメモリのデータバックアップ動作中の状態を示す図である。

【図１１】図９に示した実施例のデータバックアップ時の動作を示すフローチャートである。

【図１２】図９に示したディスクドライブを用いた本発明によるディスクアレイ装置のさらに他の実施例を示すブロック図である。

【図１３】図１２における各ディスクドライブでのデータ書込み状態を示す図である。

【図１４】図１２に示した実施例のデータ書込み，データ読出し動作を示すタイミング図である。

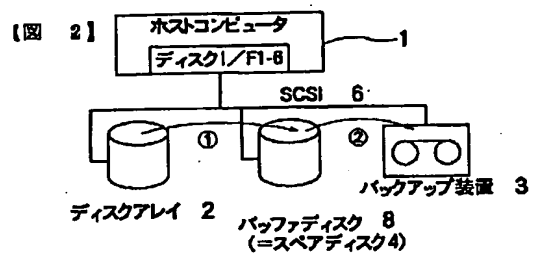
【図１５】図１２に示した実施例でのデータ回復動作を示すタイミング図である。

【図１６】図１２に示した実施例でのデータバックアップ時の動作を示すタイミング図である。

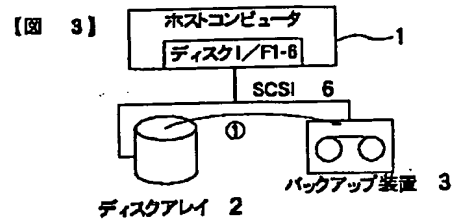
【符号の説明】

１　ホストコンピュータ
２　ディスクアレイコントローラ
３　バックアップ装置
４　スペアディスクドライブ
５　データ／パリティディスクドライブ
６　ホスト側ＳＣＳＩバス
７　ドライブ側ＳＣＳＩバス
１０　メモリ制御部
１１　データメモリ
１２　フォーマット制御部
１５　ディスク
１６　書込禁止領域判定部
２０　ホストコンピュータ
２１　ディスクアレイ装置
２３　データ分配制御部
２４　バッファメモリ
２５　パリティ生成部
２６　データ回復部
２８　磁気テープ装置
２９ａ～２９ｅ　ＳＣＳＩ制御部
３０ａ～３０ｅ，３１ａ～３１ｅ　ディスクドライブ
３３，３４　ＳＣＳＩ制御部

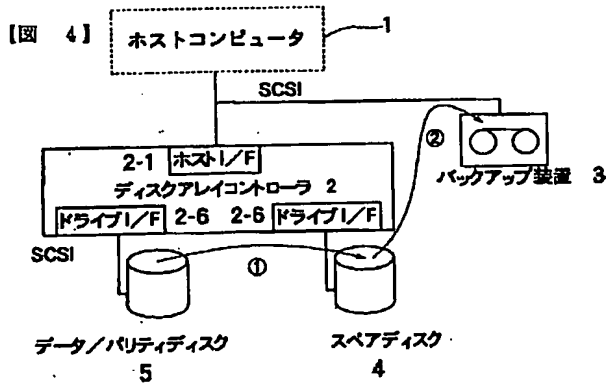
【図２】



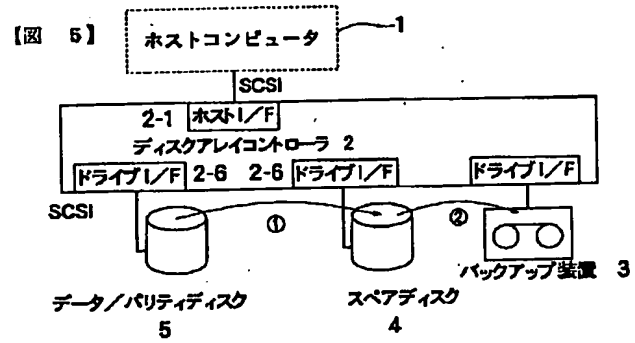
【図３】



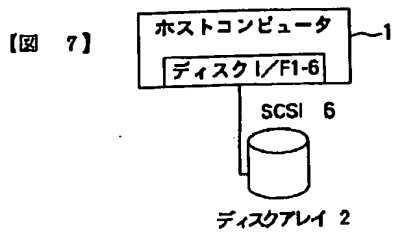
【図４】



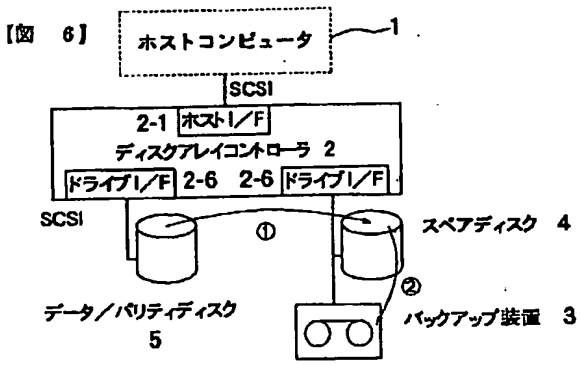
【図５】



【図７】

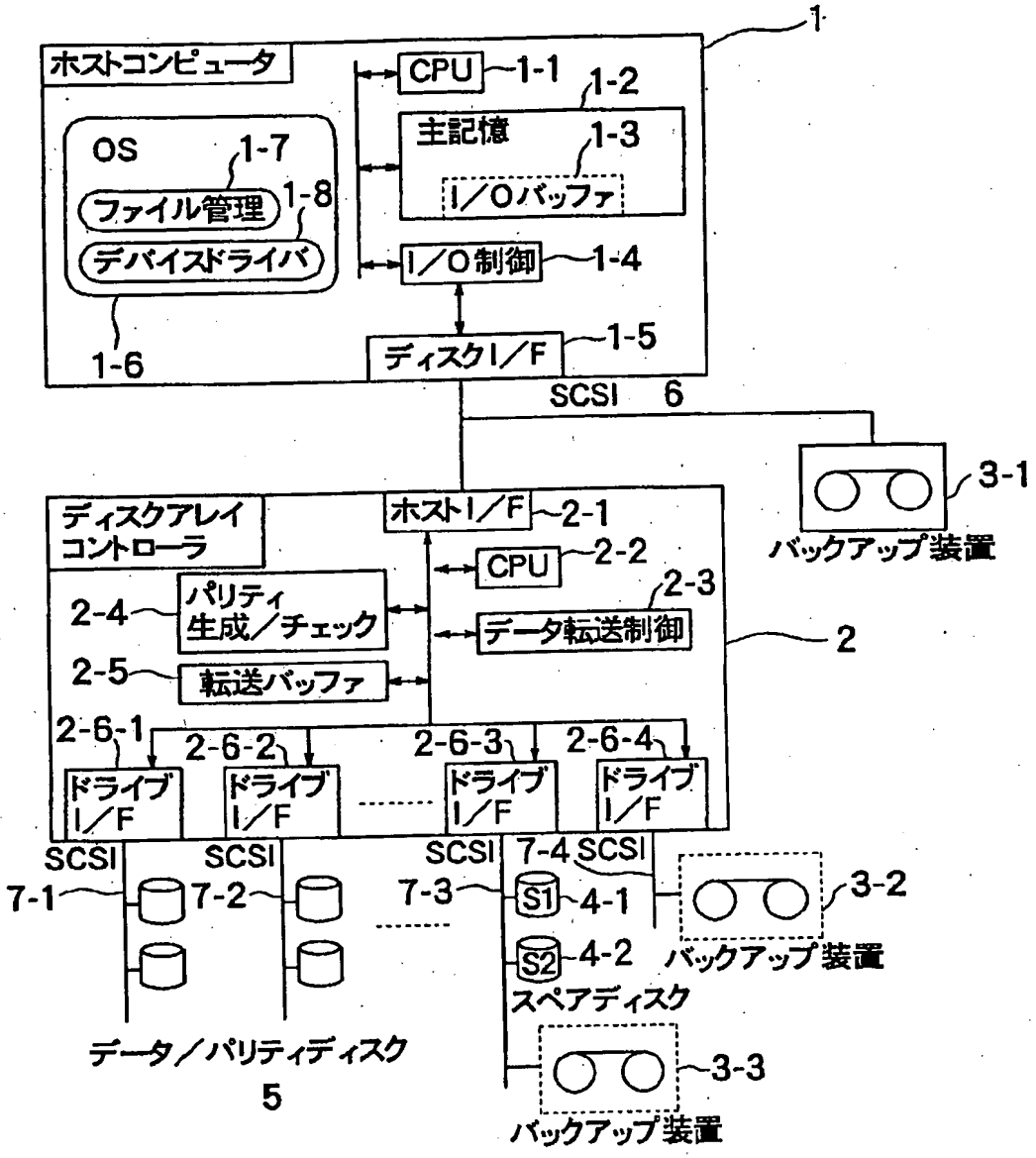


【図６】

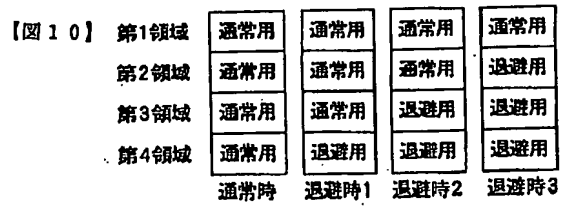


【図１】

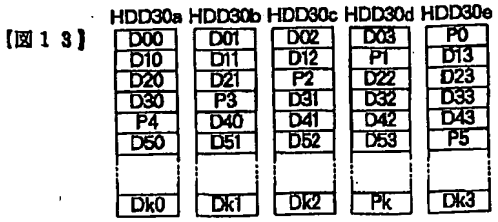
【図　１】


1
ホストコンピュータ
CPU
1-1
1-2
主記憶
1-3
I／Oバッファ
OS
1-7
ファイル管理
1-8
デバイスドライバ
I／O制御
1-4
1-6
ディスクI／F
1-5
SCSI　6
3-1
バックアップ装置
ディスクアレイ
コントローラ
ホストI／F
2-1
CPU
2-2
2-3
2-4
パリティ
生成／チェック
データ転送制御
2
2-5
転送バッファ
2-6-1
2-6-2
2-6-3
2-6-4
ドライブ
I／F
SCSI
7-1
7-2
7-3
7-4
S1
4-1
3-2
バックアップ装置
S2
4-2
スペアディスク
データ／パリティディスク
5
3-3
バックアップ装置

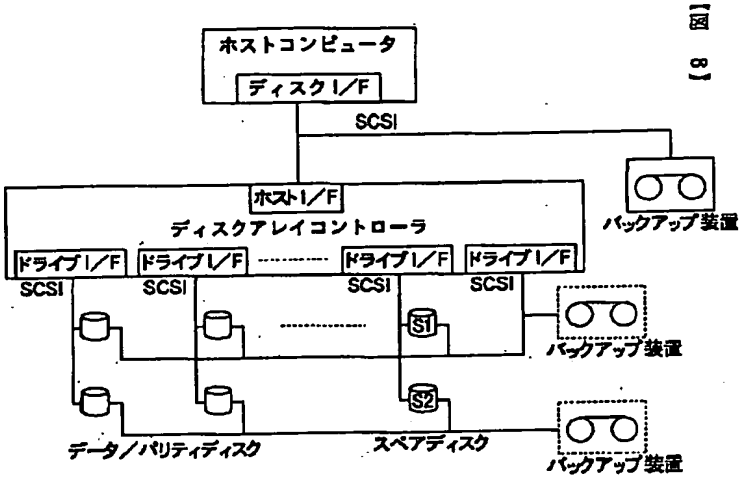
【図１０】


【図１０】
第1領域
第2領域
第3領域
第4領域
通常用
退避用
通常時
退避時1
退避時2
退避時3

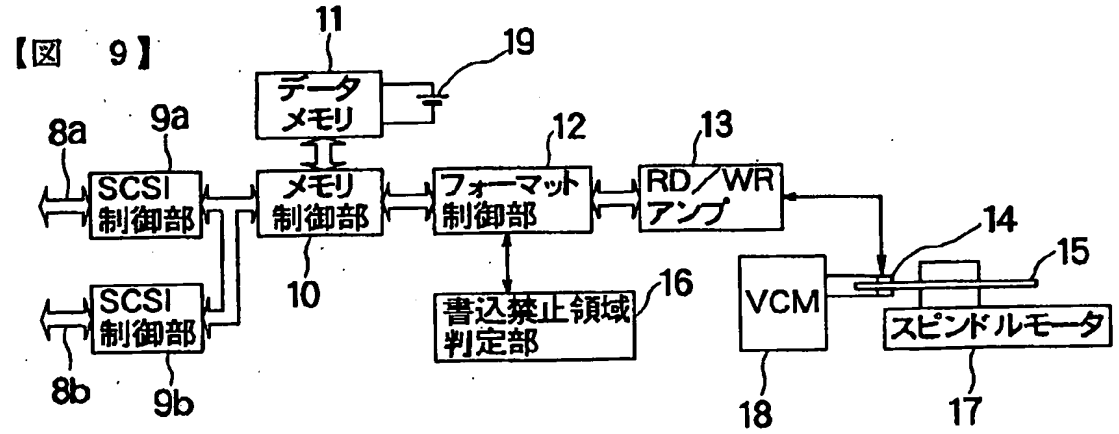
【図１３】


【図１３】
HDD30a
HDD30b
HDD30c
HDD30d
HDD30e
D00
D01
D02
D03
P0
D10
D11
D12
P1
D13
D20
D21
P2
D22
D23
D30
P3
D31
D32
D33
P4
D40
D41
D42
D43
D50
D51
D52
D53
P5
Dk0
Dk1
Dk2
Pk
Dk3

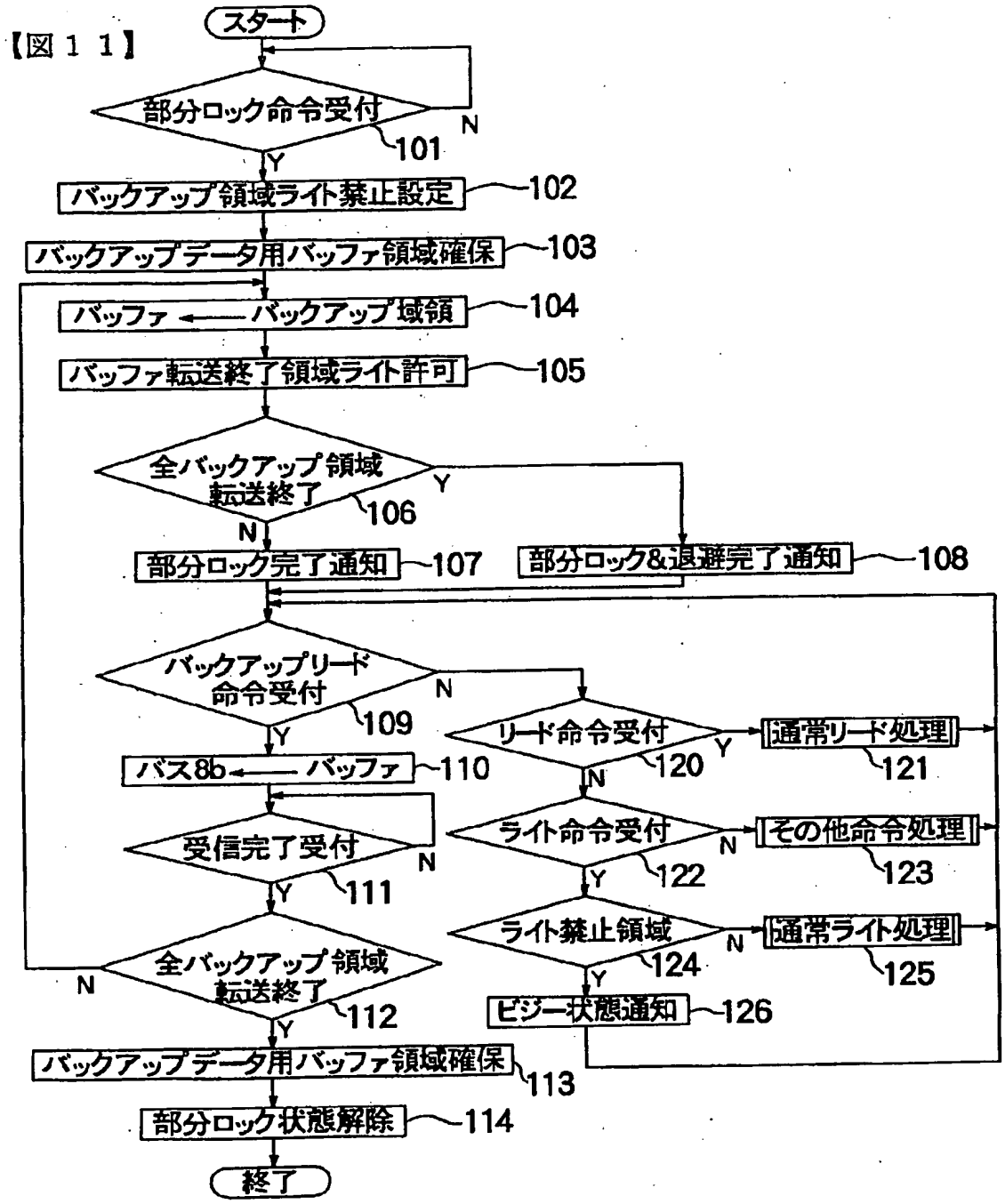
【図８】


【図８】
ホストコンピュータ
ディスクI/F
SCSI
ホストI/F
ディスクアレイコントローラ
ドライブI/F
ドライブI/F
ドライブI/F
ドライブI/F
SCSI
SCSI
SCSI
SCSI
S1
S2
バックアップ装置
バックアップ装置
バックアップ装置
データ/パリティディスク
スペアディスク

【図９】

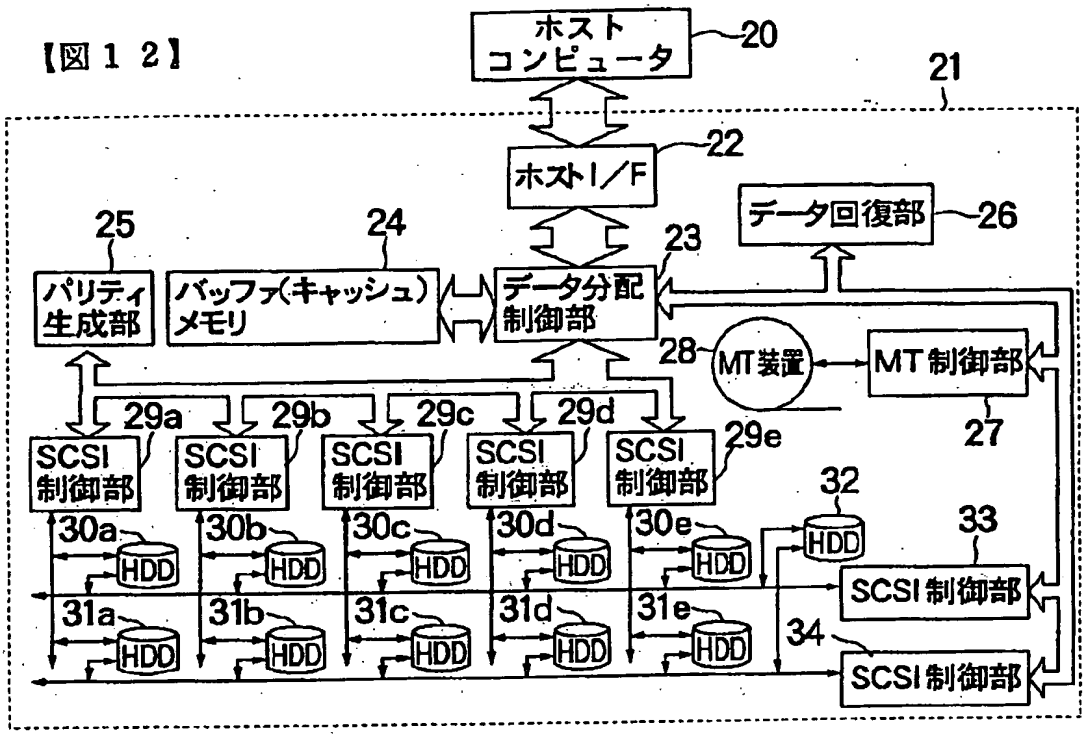

【図　9】
11
19
データ
メモリ
8a
9a
SCSI
制御部
メモリ
制御部
フォーマット
制御部
12
13
RD／WR
アンプ
14
15
VCM
スピンドルモータ
SCSI
制御部
10
書込禁止領域
判定部
16
8b
9b
18
17

【図１１】

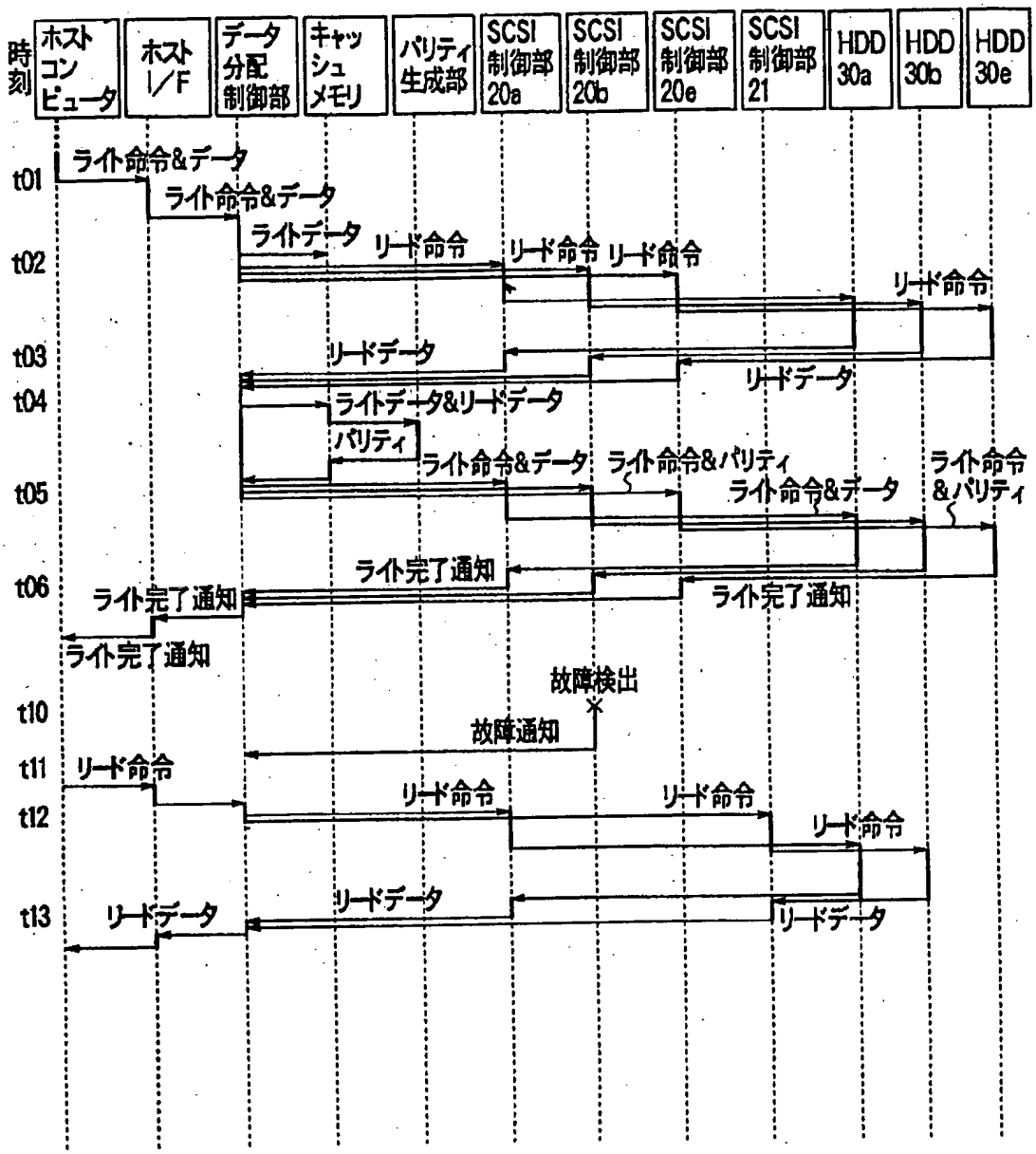

【図１１】
スタート
部分ロック命令受付
101
N
Y
バックアップ領域ライト禁止設定
102
バックアップデータ用バッファ領域確保
103
バッファ ← バックアップ域領
104
バッファ転送終了領域ライト許可
105
全バックアップ領域転送終了
106
Y
N
部分ロック完了通知
107
部分ロック&退避完了通知
108
バックアップリード命令受付
109
N
Y
バス8b ← バッファ
110
受信完了受付
111
N
Y
全バックアップ領域転送終了
112
N
Y
バックアップデータ用バッファ領域確保
113
部分ロック状態解除
114
終了
リード命令受付
120
Y
N
通常リード処理
121
ライト命令受付
122
N
Y
その他命令処理
123
ライト禁止領域
124
N
Y
通常ライト処理
125
ビジー状態通知
126

【図１２】

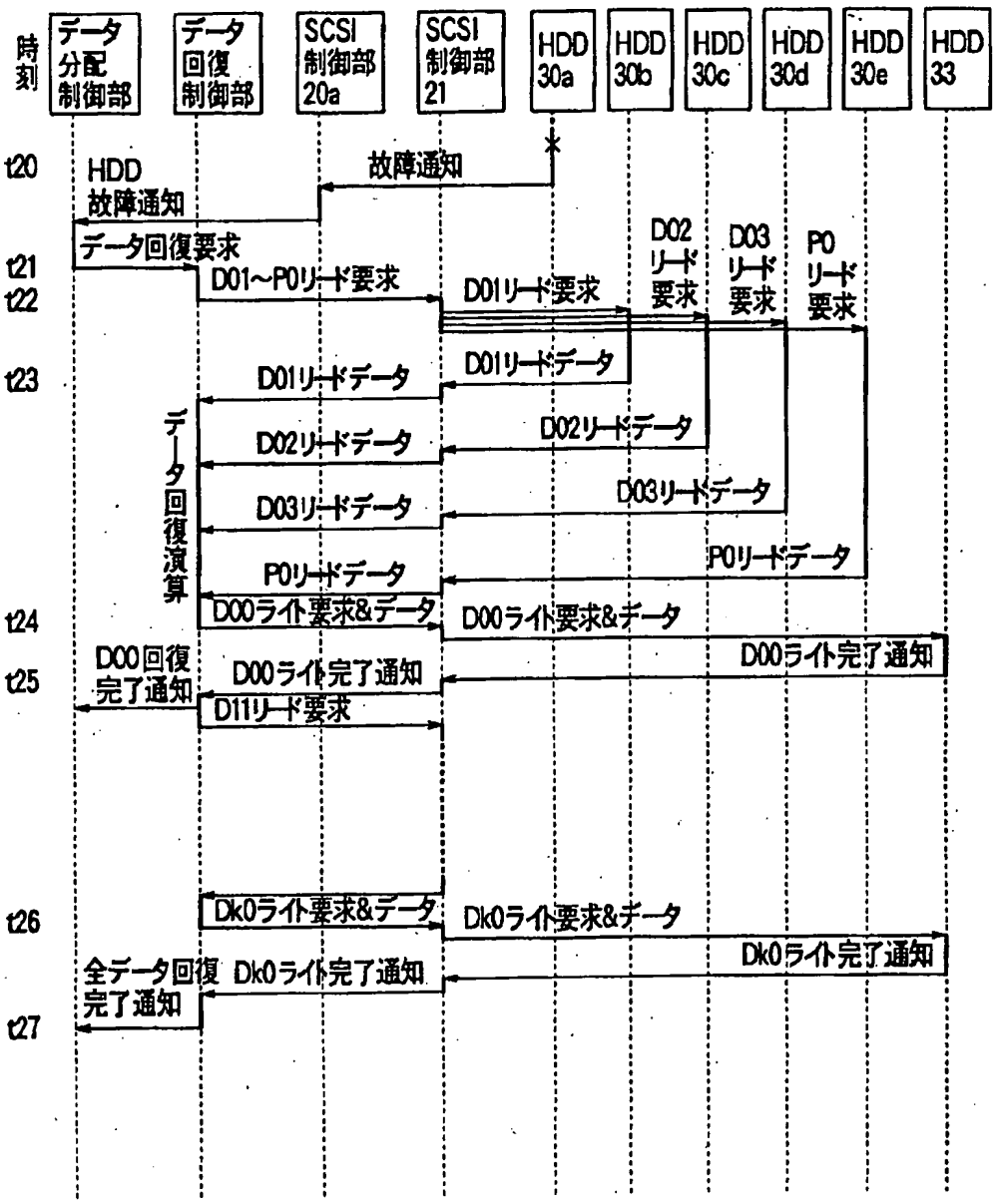
【図１２】


ホスト
コンピュータ
20
21
22
ホストI／F
データ回復部
26
25
24
23
パリティ
生成部
バッファ(キャッシュ)
メモリ
データ分配
制御部
28
MT装置
MT制御部
27
29a
29b
29c
29d
29e
SCSI
制御部
32
33
30a
30b
30c
30d
30e
HDD
SCSI制御部
31a
31b
31c
31d
31e
34

【図１４】

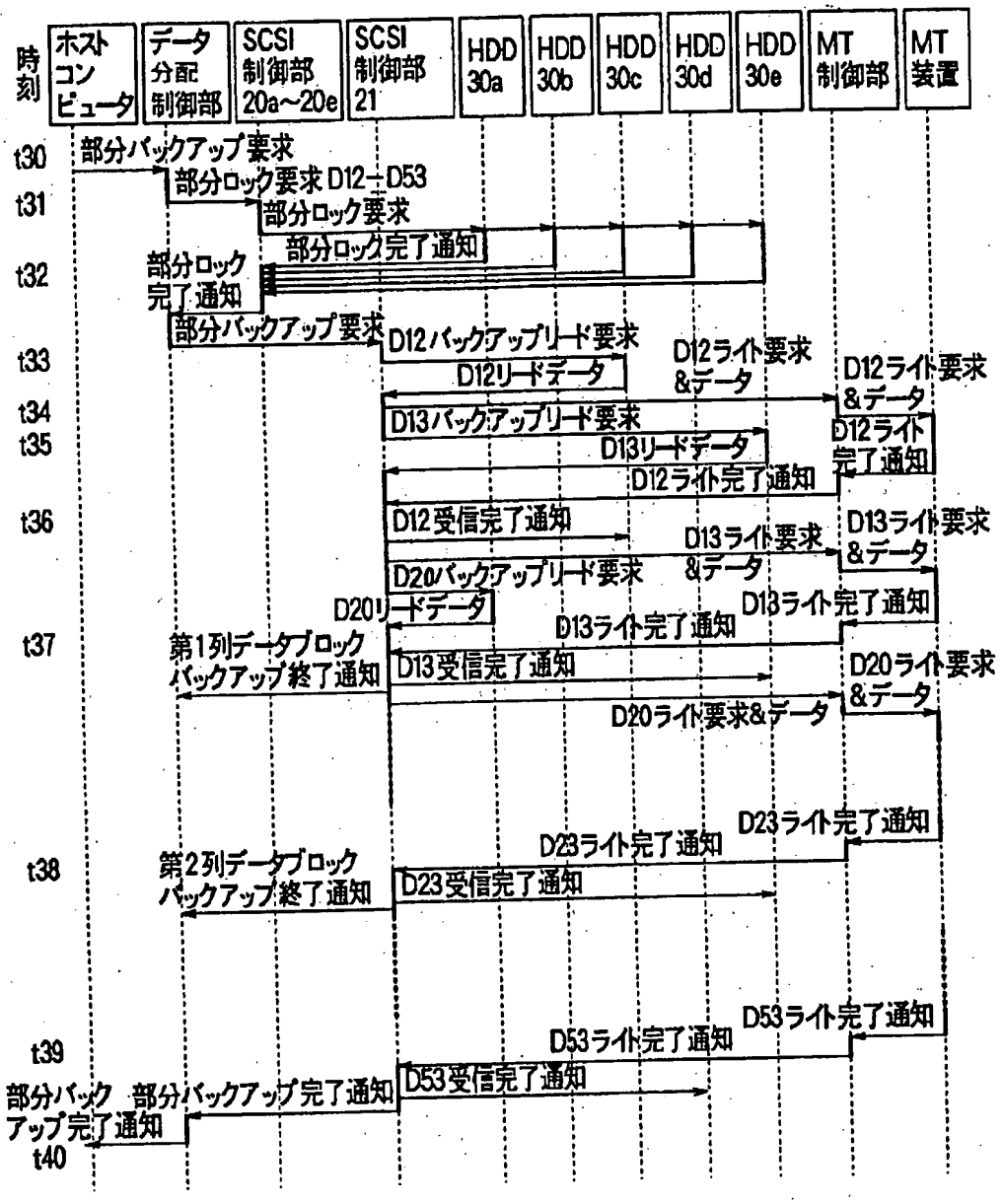

時刻
ホストコンピュータ
ホストI/F
データ分配制御部
キャッシュメモリ
パリティ生成部
SCSI制御部20a
SCSI制御部20b
SCSI制御部20e
SCSI制御部21
HDD 30a
HDD 30b
HDD 30e
t01
t02
t03
t04
t05
t06
t10
t11
t12
t13
ライト命令&データ
ライトデータ
リード命令
リードデータ
ライトデータ&リードデータ
パリティ
ライト命令&パリティ
ライト完了通知
故障検出
故障通知

【図１５】


【図１５】
時刻
データ分配制御部
データ回復制御部
SCSI制御部20a
SCSI制御部21
HDD 30a
HDD 30b
HDD 30c
HDD 30d
HDD 30e
HDD 33
t20
HDD故障通知
故障通知
データ回復要求
t21
D01～P0リード要求
t22
D01リード要求
D02リード要求
D03リード要求
P0リード要求
t23
D01リードデータ
D01リードデータ
D02リードデータ
D02リードデータ
D03リードデータ
D03リードデータ
P0リードデータ
P0リードデータ
データ回復演算
t24
D00ライト要求&データ
D00ライト要求&データ
D00ライト完了通知
t25
D00回復完了通知
D00ライト完了通知
D11リード要求
t26
Dk0ライト要求&データ
Dk0ライト要求&データ
Dk0ライト完了通知
全データ回復完了通知
Dk0ライト完了通知
t27

【図１６】

【図１６】



フロントページの続き

(72)発明者　本田　聖志
神奈川県横浜市戸塚区吉田町292番地　株式会社日立製作所マイクロエレクトロニクス機器開発研究所内

(72)発明者　松並　直人
神奈川県横浜市戸塚区吉田町292番地　株式会社日立製作所マイクロエレクトロニクス機器開発研究所内

(72)発明者　宮沢　章一
神奈川県横浜市戸塚区吉田町292番地　株式会社日立製作所マイクロエレクトロニクス機器開発研究所内

(72)発明者　磯野　聡一
神奈川県横浜市戸塚区吉田町292番地　株式会社日立製作所マイクロエレクトロニクス機器開発研究所内